Easy-to-use semantic search of pharmacological data

Guillermo Vega-Gorgojo[1] and Laura Slaughter[2]

[1] Department of Informatics, University of Oslo, Norway guiveg@ifi.uio.no
[2] Oslo University Hospital, Norway laura.slaughter@gmail.com

Abstract. Patient safety and treatment effectiveness can be improved by introducing pharmacogenomic testing into current clinical practice. LOD resources such as DrugBank and SIDER are readily available to be used but are not integrated and cannot be easily exploited by clinical health workers. To overcome these limitations, we have set up a novel pharmacological search facility that combines data from multiple RDF sources and uses PepeSearch to formulate queries. We demonstrate this search system that is currently being tested with clinicians as a means for exploring the knowledge contained in these databases, to create flexible queries and to support decision-making.

1 Introduction and Motivation

Access to evidence-based pharmaceutical knowledge with accompanying genetic information must be incorporated into the systems currently used in healthcare. By introducing pharmacogenomic testing into current clinical practice both patient safety and treatment effectiveness can be improved. However, Taber & Dickinson [1] have shown that physicians lack knowledge about the topic and need educational as well as informational resources. Making this information available has been studied related to the design of Computerized Provider Order Entry systems (CPOE) having context-sensitive information combined with alerts [2]. Essential work by Romagnoli et al. [3] has focused on information needs of hospital pharmacists, who have the complex job of handling difficult patient cases and medication reconciliation tasks.

Physicians and hospital pharmacists can benefit from the use of a pharmacological knowledge base composed of Linked Open Data (LOD) resources and additional information from FDA labelling. RDF datasets, such as DrugBank and the Side Effect Resource SIDER are readily available to be used, but unfortunately, (1) databases are not integrated, and (2) search facilities are inadequate. To overcome these limitations we are working on providing an easy-to-use search facility of pharmacological data. In this work, we have chosen to focus on physicians and pharmacists' information needs, i.e. for clinical prescription. Other potential cases such as drug discovery (see [4] and [5]) will expand this work in future efforts. Using input from [3], the current work focuses on integration and searching multiple open drug Linked Data sources for general information needs proposed by at least 25% of pharmacists in this study. These information needs are listed in Fig. 1.

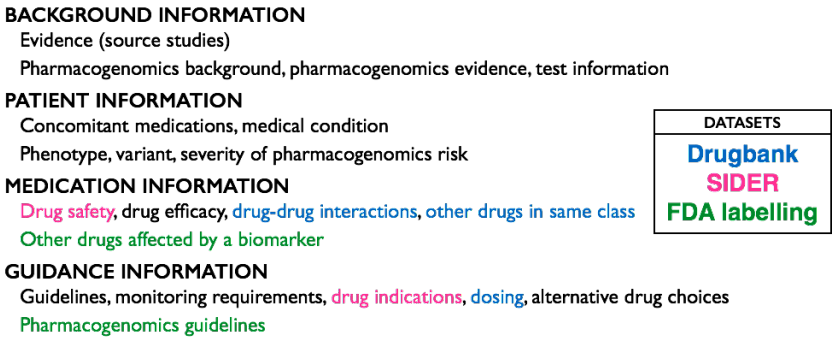

Fig. 1. Pharmacists' information needs extracted from [3] and dataset coverage.

2 A Novel Pharmacological Search Facility

We aim to create a pharmacological resource based on existing LOD drug databases that can be used to fulfil the information needs of pharmacists and physicians. As a foundation we have employed DrugBank, a comprehensive repository of drug, drug-target and drug action information based on extensive literature surveys and curated by experts [6]. DrugBank is one of the most popular resources in the pharmacological domain due to its wide coverage and to the links to other well-known databases such as ATC, ChEBI, PubChem and KEGG. More importantly, DrugBank provides relevant medication information to pharmacists (see Fig. 1) and Bio2RDF already offers an RDF version of DrugBank.[1]

In addition to DrugBank, we have employed the Side Effect Resource SIDER [7] to obtain information about drug safety and drug indications (see Fig. 1). SIDER extracts indications and adverse drug reactions (ADRs) of marketed medicines from package inserts and the biomedical literature. Bio2RDF also exposes an RDF version of SIDER,[2] thus facilitating the integration with DrugBank. In this regard, the two datasets employ different URIs, so we found matches based on the drug name. We generated 1179 mapping triples to link the drugs in the two databases.

Finally, our main goal was to include pharmacogenomic information, given its importance to reduce the risks of adverse events and to improve the effectiveness to treatments [3]. We have employed the list of FDA-approved drugs with pharmacogenomic information in their labeling, including specific actions to be taken based on the biomarker information [8]. Since this resource is not available as LOD, we have translated this information into RDF and linked it to the drug URIs in DrugBank. As a result of this work, we have set up a triple store that integrates DrugBank, SIDER and FDA's pharmacogenomic data.

As pharmacists and physicians need a search tool that allows them to easily express their information needs without requiring knowledge of SPARQL or

[1] `http://download.bio2rdf.org/release/3/drugbank/drugbank.html`
[2] `http://download.bio2rdf.org/release/3/sider/sider.html`

RDF, we have employed PepeSearch [9] for this purpose. PepeSearch is a portable form-based search interface for SPARQL endpoints that allows the searcher to set multiple constraints on any of the classes in the target dataset. For an arbitrary RDF class, PepeSearch creates a form block in which datatype properties are mapped to widget elements, e.g. text boxes for string literals or slide ranges for integers. In this way, a searcher can easily set restrictions on a class by manipulating the visual elements of the block form. Beyond refining a single class, PepeSearch allows the formulation of queries that involve multiple classes; the user interface will include new form blocks for each class connected with an object property to the selected class.

We have thus set up a PepeSearch instance for querying our pharmacological triple store at `http://sws.ifi.uio.no/project/drugsearch/`. Note that we have pruned some of the classes employed in Drugbank and SIDER (such as carrier information and polarizability) for the sake of simplicity. Our criterion has been to only cover the information needs identified in [3] (see Fig. 1).

3 Demonstration

To illustrate the operation of the resulting search facility, we will employ the following information need: *"obtain the drugs indicated for myocardial infarction that elicit variable responses for patients with biomarker CYP2C19"*. It is inspired in the pharmacogenomic study by Taber & Dickinson [1] and requires the usage of the three data sources of our pharmaceutical database: drugs (DrugBank), indications (SIDER) and pharmacogenomic biomarkers (FDA). Specifically, the interaction with PepeSearch can be the following:

1. PepeSearch presents a list of the top classes available – see Fig. 2(a).
2. The searcher selects "Drug".
3. PepeSearch presents a form block for the "Drug" class and a list of collapsibles corresponding to related classes: "Indication", "Biomarker"...
4. The searcher sets the restrictions required for this search task: she expands the "Indication" collapsible and fills in "myocardial infarction"; then she expands the "Biomarker" collapsible and fills in "CYP2C19" – see Fig. 2(b).
5. The searcher pushes the "Get results" button at the top right corner.
6. PepeSearch generates a SPARQL query that is sent to the endpoint.
7. PepeSearch prepares a tabular representation of the results – see Fig. 2(c).
8. The searcher can navigate through the results by following the links in the table, e.g. to obtain additional information about Clopidogrel.

4 Discussion and Future Work

We have introduced a search system for integrated LOD pharmaceutical data which is a promising resource that can provide clinicians a means for exploring the knowledge contained in these databases, to create flexible queries and to

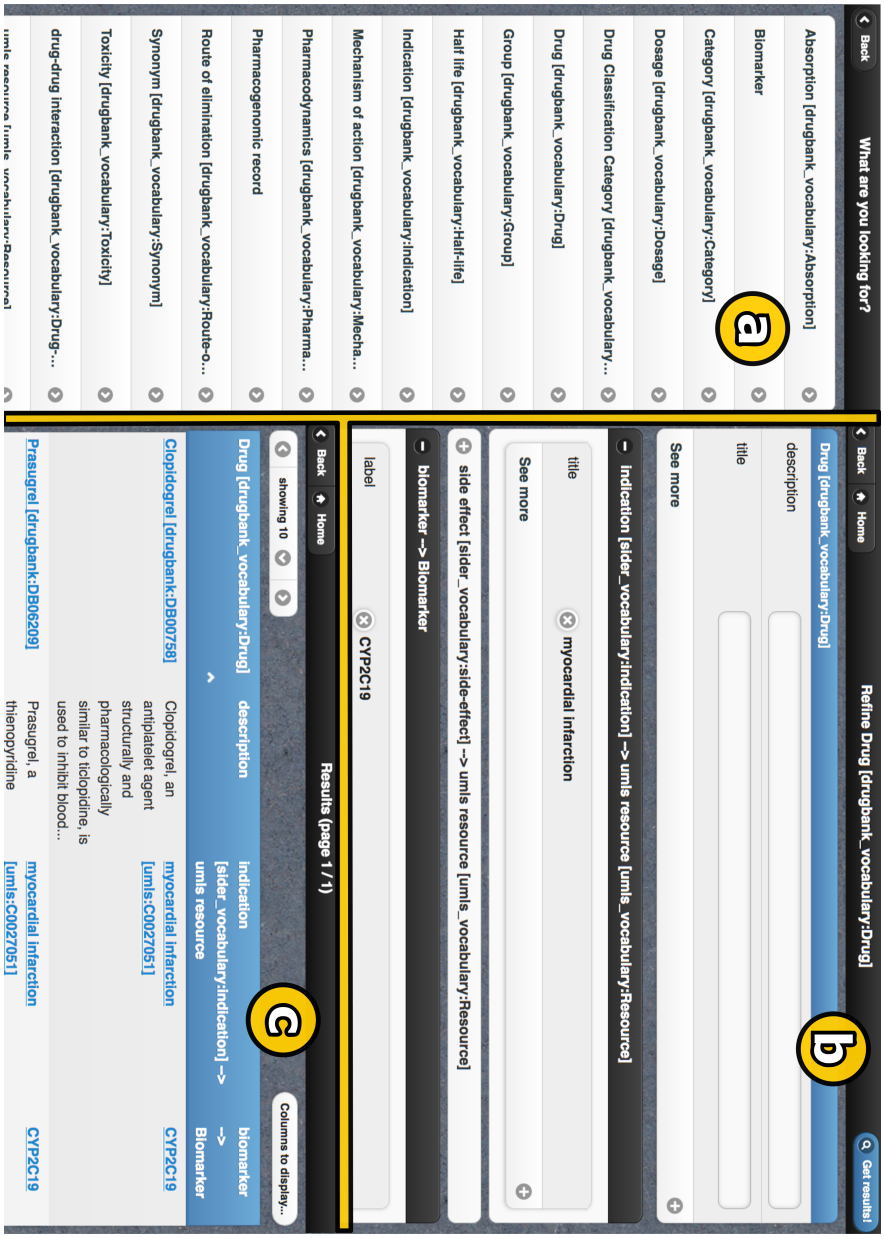

Fig. 2. Snapshots of PepeSearch.

support decision-making. The benefit to this search tool includes cost-effective reuse of existing datasets, a simple form-based query interface, and a means to express information needs within multiple fields for more precise results.

The information needs for hospital pharmacists and clinicians for pharmacogenomics are different from researchers' needs. The majority of datasets and resources are geared towards serving research needs. Callahan et al. [10] discussed difficulties with searching and integrating biological data in the open-source Bio2RDF project. We have taken the first step towards reuse of datasets for the "average hospital clinician". Our current research involves user testing of the clinicians' ability to express their information needs using our search facility. Further future steps include the use of PharmaGKB [11] as a key resource to implement pharmacogenomics in real-world clinical practice.

Acknowledgements

This work has been funded by the BIGMED (IKT 259055), HealthInsight (NFR 247784/O70), Optique (FP7 GA 318338), and BYTE (FP7 GA 619551) projects.

References

1. Taber, K.A.J., Dickinson, B.D.: Pharmacogenomic knowledge gaps and educational resource needs among physicians in selected specialties. Pharmacogenomics and Personalized Medicine **7** (2014) 145–162
2. Devine, E.B., Lee, C.J., Overby, C.L., Abernethy, N., McCune, J., Smith, J.W., Tarczy-Hornoch, P.: Usability evaluation of pharmacogenomics clinical decision support aids and clinical knowledge resources in a computerized provider order entry system: a mixed methods approach. International Journal of Medical Informatics **83**(7) (2014) 473–483
3. Romagnoli, K.M., Boyce, R.D., Empey, P.E., Adams, S., Hochheiser, H.: Bringing clinical pharmacogenomics information to pharmacists. International Journal of Medical Informatics **86** (2016) 54–61
4. Kim, R.S., Goossens, N., Hoshida, Y.: Use of big data in drug development for precision medicine. Expert Review of Precision Medicine and Drug Development **1**(3) (2016) 245–253
5. Dumontier, M., Wild, D.J.: Linked data in drug discovery. IEEE Internet Computing **16**(6) (2012)
6. Law, V., Knox, C., Djoumbou, Y., Jewison, T., Guo, A.C., Liu, Y., Maciejewski, A., Arndt, D., Wilson, M., Neveu, V., et al.: Drugbank 4.0: shedding new light on drug metabolism. Nucleic Acids Research **42** (2014) 1091–1097
7. Kuhn, M., Letunic, I., Jensen, L.J., Bork, P.: The sider database of drugs and side effects. Nucleic Acids Research **44** (2015) 1075–1079
8. Food and Drug Administration: Table of pharmacogenomic biomarkers in drug labeling (2016) URL: `http://www.fda.gov/Drugs/ScienceResearch/ResearchAreas/Pharmacogenetics/ucm083378.htm`.
9. Vega-Gorgojo, G., Giese, M., Heggestøyl, S., Soylu, A., Waaler, A.: PepeSearch: Semantic data for the masses. PLOS ONE (2016)
10. Callahan, A., Cruz-Toledo, J., Dumontier, M.: Ontology-based querying with Bio2RDF's linked open data. Journal of biomedical semantics **4**(Suppl 1) (2013)
11. Whirl-Carrillo, M., McDonagh, E., Hebert, J., Gong, L., Sangkuhl, K., Thorn, C., Altman, R., Klein, T.E.: Pharmacogenomics knowledge for personalized medicine. Clinical Pharmacology and Therapeutics **92**(4) (2012) 414